2013 IEEE International Conference on Computer Vision

Drosophila Embryo Stage Annotation using Label Propagation

Tomáš Kazmar[1,2] Evgeny Z. Kvon[1] Alexander Stark[1] Christoph H. Lampert[2]

[1]Research Institute of Molecular Pathology (IMP)
Dr. Bohr-Gasse 7, Vienna, Austria
{kazmar, kvon, stark}@imp.ac.at

[2]Institute of Science and Technology Austria
Am Campus 1, Klosterneuburg, Austria
chl@ist.ac.at

Abstract

In this work we propose a system for automatic classification of Drosophila embryos into developmental stages. While the system is designed to solve an actual problem in biological research, we believe that the principle underlying it is interesting not only for biologists, but also for researchers in computer vision.

The main idea is to combine two orthogonal sources of information: one is a classifier trained on strongly invariant features, which makes it applicable to images of very different conditions, but also leads to rather noisy predictions. The other is a label propagation step based on a more powerful similarity measure that however is only consistent within specific subsets of the data at a time.

In our biological setup, the information sources are the shape and the staining patterns of embryo images. We show experimentally that while neither of the methods can be used by itself to achieve satisfactory results, their combination achieves prediction quality comparable to human performance.

1. Introduction

In biological research, high-throughput screening has become a common technique for producing microscopic data. Thanks to the automation of microscopes and robotic sample feeders, large volumes of image data can be produced almost completely automatically. As a consequence, image annotation has become the main bottleneck in biological image processing. Classification tasks, such as distinguishing between different developmental stages, and regression tasks, such as estimating an organism's pose in an image, are still done almost exclusively by manual inspection and annotation.

In this work we tackle the following problem: given an image containing many embryos of the fruit fly *Drosophila melanogaster*, identify for each embryo which out of six developmental stage groups it is in, see Figure 1.

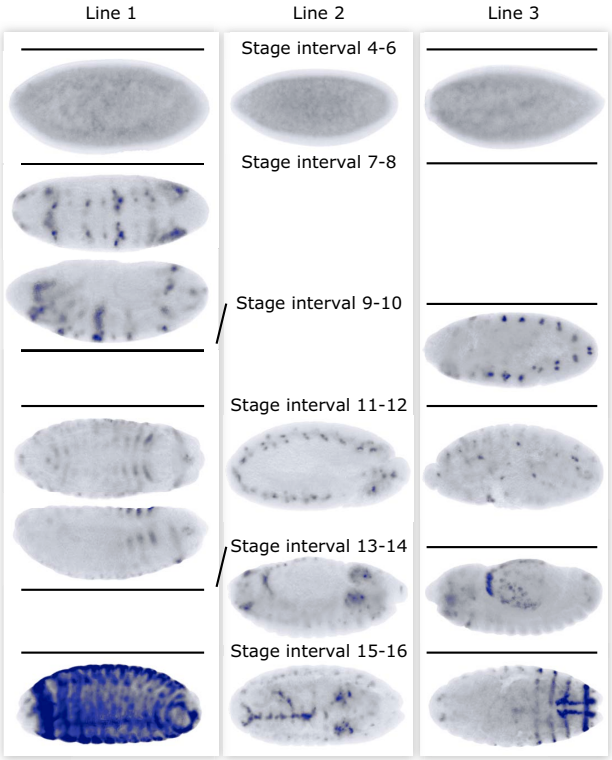

Figure 1. Examples of embryo images of three differently genetically modified Drosophila lines (left, middle, right). The modifications cause different appearances, which change as the embryos develop (top-to-bottom). The pattern is often indicative of stage inside the line, even though it can vary due to different pose of the embryo (stages 7–8 and 11–12 of Line 1). Across different lines the pattern for the same stage usually differs significantly (compare embryos of stage 15–16). Images of stage 4–6 show no activity; in other stages empty space means no activity pattern.

In the rest of this section, we provide the biological background and motivation for this task. We then explain the technical contribution in Section 2, our experimental evaluation in Section 3 and we end with a discussion in Section 4.

1550-5499/13 $31.00 © 2013 IEEE
DOI 10.1109/ICCV.2013.139

Biological motivation. Our motivation comes from a large-scale biological study aimed at deciphering the function of non-coding regions in the genome. For this, many small genomic DNA fragments are tested for their gene regulatory function [9]. Each of the fragments to be tested is integrated in the Drosophila genome together with a reporter gene. Subsequently, the resulting embryo is allowed to develop up to a duration of 24 hours. If the fragment has a regulatory function and becomes active during development, the reporter gene is co-activated, and can be visualized by *in-situ hybridization* [13] giving rise to a visible pattern inside the otherwise transparent embryo. This pattern is characteristic for the function of the DNA fragment. It typically corresponds to a well-defined anatomical structure, such as the developing brain, or muscle tissue, and it is visible only during a specific time interval in the organism's development, see Figure 1.

Since all of the processes described above are error-prone, many measurements are required before a complete picture of the function of the DNA fragment can be formed. Therefore, hundreds of genetically identical embryos are created at the same time, mounted on a microscopic slide and imaged simultaneously using a large whole-slide microscopic setup (for a sample image, see the supplemental material). An advantage of this process is that embryos of all developmental stages occur within a single image, making it unnecessary to repeat the time-consuming and costly step of embryo preparation and image acquisition multiple times for different stages. A disadvantage is that in order to interpret the patterns and identify the function of the respective DNA fragment, the hundreds of embryos need to be separated into their respective developmental stages before a further analysis can be performed. This is necessary for every of the thousands of DNA fragments to be tested, resulting in the need to classify several million embryo images. Doing so manually would require a trained experts many years. However, the task cannot simply be outsourced to a crowdsourcing platform, such as Mechanical Turk, since the differences between stages are subtle [3], and therefore a certain amount of experience and background knowledge is required to achieve the necessary high accuracy. Consequently, a computer vision system that can automatically or semi-automatically provide stage annotation for the embryo images is a crucial element for the success of the project.

Related work. With the advancement of computer vision techniques for biological applications, also the task of analyzing Drosophila embryo images has received a certain amount of attention. One line of research focusses on the unsupervised extraction of expression patterns and appearance-based clustering. This includes the popular BEST [8] and SPEX^2 methods [14]. These techniques can identify groups of similar patterns from which one can, *e.g.*, infer prospective gene interaction networks [15]. However, they do not provide stage annotation, as we require.

Early techniques relied on generative probabilistic models, in particular Gaussian mixture models [12, 19]. More recently, other learning techniques such as SVM [18], multi-instance learning [11] and bag-of-visual-word representations [7], or random walks [2] have been explored. However, these approaches are also not directly applicable to our situation: first, they are focussed on images in a specific format, as it is present in the popular Berkeley Drosophila Genome project (BDGP) database [16, 17]. This database consists of manually selected embryos with a clearly visible pattern and per-embryo pose annotation. Neither of these properties can be ensured in a large-scale automatic screening. Second, and more fundamentally, the existing classification methods based their decision on the visual patterns in the embryo. Since these patterns change fundamentally between the different tested DNA fragments, applying the above techniques would require a separate training step for each slide. In particular, this includes creating a separate training set per slide, thereby defeating the purpose of having an automatic system to start with.

Our approach and contribution. In this work, we introduce a system for automatic staging of Drosophila embryos that leverages the information in expression patterns without requiring training data specific to any of these patterns. The main idea is to combine two orthogonal sources of information. A *base classifier* for staging is trained using only the contour (shape) of the embryo. This requires only modest amounts of training data, since the embryo contour is not affected by the genetic modifications and therefore the same classifier can be applied to all embryo images in all slides at test time. Within each slide, the resulting predictions are improved and robustified using *label propagation*, with a similarity graph that is obtained from the similarity between expression patterns (appearance). This is possible because these patterns are consistent within each slide, *i.e.*, the patterns are different between stages yet identical within a stage. It is only between different slides that they can change arbitrarily as embryos in different slides have different mutations.

Note that for both, the classifier training and the label propagation, we rely on existing techniques. Our contribution lies not in a new objective or algorithm but in the observation that in our situation shape and appearance are truly orthogonal sources of information, and in proposing a way for combining them such that the weak classification signal of the former is strengthened by the only locally consistent similarity measure of the latter.

We believe, however, that this insight will be of interest not only to researchers working on the specific biological task at hand, but that it has application also in other com-

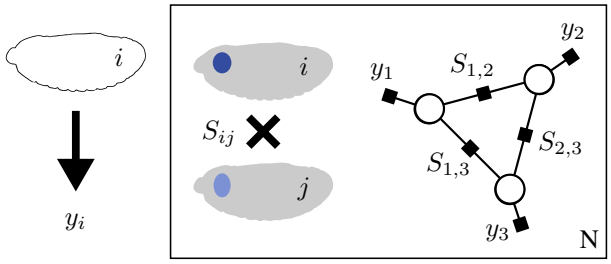

Figure 2. Scheme of the method. A base classifier uses a general source of information (embryo shape) and therefore it is trained only once, whereas the label propagation works always with a set of embryos sharing the same genotype and can take advantage of the genotype-specific features (embryo appearance).

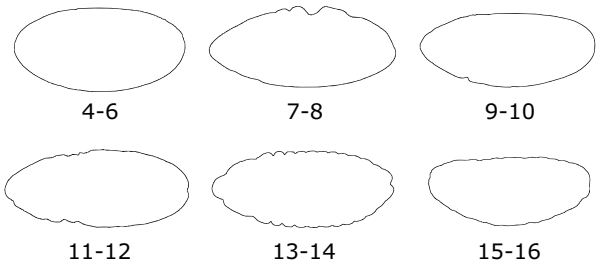

Figure 3. Examples of contours of embryos of different stages. Contours of some of the stages (4–6, 9–10) are very smooth, whereas for other stages there are easily recognizable markers reflecting the changes of embryo morphology, like the appearing segments for stages 11–12, and 13–14. Some of these markers are best visible only in some orientations (side view in the case of stage 7–8).

puter vision applications where at test time an additional degree of homogeneity is available compared to the training time. Examples could be writer-specific handwriting recognition or object recognition in personal photo collections.

2. Automatic stage annotation

Our system for automatic annotation of embryo developmental stages has three main components: 1) a shape-based classifier that assigns a probability estimate for being in each of the stage classes to each individual embryo image, 2) an appearance-based similarity measure between embryo images, which we use to form a weighted neighborhood graph between embryos within a slide, and 3) a label propagation step that combines evidence of the multiple predictions within the slide, thereby improving the overall classification accuracy (for a scheme see Figure 2). In this section, we will describe each of the individual components in more detail.

2.1. Shape-based stage classification

One way to distinguish between different developmental stages is based on the embryos' shape [3], in particular their contour (Figure 3). Compared to the appearance of the embryo, the contour shape is not affected by the genetic modifications we induce. Therefore, using a shape feature we can build a classifier that is trained just once off-line and then applied to every later embryo image, regardless of which DNA fragment had been inserted into the genome.

On a global scale, the embryo contour is always close to elliptical. We make use of this fact by automatically extracting individual embryos from the slide image and rotating them such that their main axis is aligned horizontally. On finer scales, embryos of different stages differ in short characteristic regions that occur in different locations and at different times during the development. To capture these in a feature vector, we first represent the embryo outline by a chaincode [5], such that any short substring of the code corresponds to a short boundary segment. We then follow a bag-of-visual-words approach [4], as has proven successful for many other visual categorization tasks.

From a training set of embryo contours we first perform *term frequency/ inverse document frequency (tfidf)* weighting of the exacted substrings, with developmental stages taking the role of different documents. This way, we ensure that frequent but non-informative segments are suppressed, and segments from all stages are treated with roughly equal importance. From the highest ranked substrings, $s_1, \ldots, s_n$ we create a dictionary of representative chaincodes using kernel vector quantization (KVQ) [10]. This requires solving a linear program

$$w^* = \underset{w \in [0,1]^n}{\arg\min} \sum_{i=1}^{n} w_i \quad \text{sb.t.} \quad \sum_{i=1}^{n} a_{ij} w_j \geq 1, \qquad (1)$$

where $a_{ij} = 1$, if the Hamming distance between substrings s_i and s_j is smaller than a chosen threshold, and $a_{ij} = 0$ otherwise. The codebook consists of all substrings that have positive weights $w_i^* > 0$. We choose this way to arrive at a representative subset over, *e.g.*, k-means clustering, because 1) k-means would requires a vector representation, whereas KVQ only requires a similarity measure (here Hamming distance) between the objects to be clustered, and 2) the only free parameter of KVQ is the threshold for defining the a_{ij}, which corresponds to the maximal radius we allow a cluster to have. This is more easily interpretable than having to specify the number of codebook entries in advance. To obtain shape features of different scale, we repeat the above procedure also on scaled-down versions of the training images.

To represent an embryo image, we extract contour segments on all scales, assigning each segment the cluster ID of its nearest codebook entry, as measured by Hamming distance. The cluster IDs are then combined into one bag-of-visual-words histogram per scale and the resulting his-

tograms of all scales are concatenated into a single feature vector.

Using this representation, we train a linear support vector machine (SVM) with Platt scaling to get a classifier with probabilistic outputs.

2.2. Breeding the training data

Collecting and annotating training data is the most time- and work-consuming stage of many visual recognition problems.

Interestingly, the situation is different for us. Since the shape-based classifier relies on features that are not affected by the tested genetic modification, we need only a single training set to get a classifier that can be applied to all future embryo images. Also, the training examples do not even have to come from genetically modified organisms. We can take any *normal* (wild-type) Drosophila line and take images of its embryos to form a training set. Most importantly, we can make use of the fact that the target classes we want to predict are developmental stages, which themselves are defined by certain intervals of how many hours the embryo has been developing. In combination, this is allows us to selectively create training examples of each stage instead of following the usual approach of collecting a large corpus of unlabeled examples to be annotated manually. All we have to do is to precisely time the interval between when the eggs start developing, and when the embryos are prepared for imaging. This is not a trivial task, but doable using existing biological techniques.

Note that this trick of breeding a training set works most efficiently for the genetically unmodified Drosophila, which can be bred in large quantities and create a large amount of healthy embryos per generation. We cannot use the same trick to get stage specific embryo images for the genetically modified organisms, since this would require many repetitions of breeding, staining and imaging for each stage of each DNA segment to be tested.

In addition to the training data obtained by breeding we also create a smaller amount of annotated images of genetically modified embryos with consistent pattern in the traditional, manual way. These serve as a validation and as a test set for the experimental evaluation we present in Section 3.

2.3. Pattern similarity

A good similarity graph is a key element for the label propagation method we want to apply. Its most important component is the similarity measure chosen. Typical measures of pattern similarity divide the embryo area into a grid or a triangulation of cells and compare the average intensities within each cell, for example by the sum of squared differences (SSD) [6]. Alternatively, it has been proposed to extract and compare SIFT descriptors on a rectangular grid [14]. Such two-dimensional similarities work well in

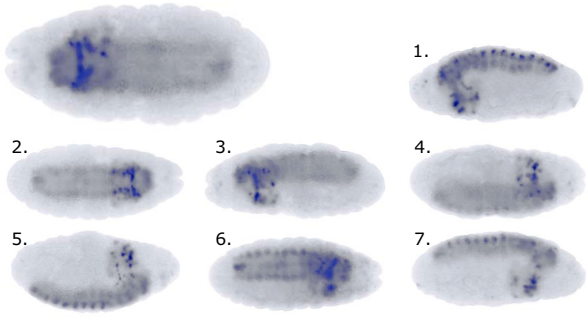

Figure 4. A stage 15–16 embryo and its 7 closest neighbors according to our measure of similarity of expression patterns. The similarity is invariant to rotation around the embryo's main axis and to in-plane rotation as well leading to all relevant embryos that should share the same stage classification being close to each other.

situations when embryo images are registered, but they are very sensitive to out-of-plane rotation, which can drastically change the appearance between embryos of the same stage, see Figure 4. We therefore adopt a similarity measure that explicitly enforces the invariance properties we know are necessary: for achieving rotational invariance along the embryo's main axis, we project the two-dimensional intensity values to the main axis of the elliptic shape by summing (the non-stained parts of the embryo are transparent, so there is no problem of occlusion). We achieve invariance to contrast variations by using the *normalized cross-correlation score* as similarity measure between such projected intensity profiles. Finally, we get a similarity between two embryo images that is also invariant to in-plane rotation by comparing the images always in two ways, once directly as described above, and once after flipping one of the images along the vertical axis, keeping the larger of the two similarity values.

Note that for a generic classification system, this similarity might not be strong enough: it is possible that the projection and the normalization remove not just geometric distortion but also relevant pattern information. In our situation, however, we found this not to be a major problem. The reason is likely that we only measure the similarity between embryo images within the same slide, where typically only few and characteristically different patterns occur.

2.4. Label propagation

At test time, given a slide with n embryo images, we form a symmetric k-NN graph of n nodes. Each edge (i, j) is assigned a weight $w_{ij} = \exp(\gamma s(x_i, x_j))$, where $s(x_i, x_j)$ is the similarity score between the images x_i and x_j, and γ is a bandwidth parameter which is found by model selection.

Since ordinary label propagation handles only binary classification, we use a one-vs-rest approach, solving one

label propagation step for each of the six stage groups we are interested in. In each case, each node i is assigned a *label preference* score y_i^l, which is the output of the shape-based classifier for this stage on the corresponding embryo image. The idea of label propagation is that the decisions of images, for which the classifier is confident (y_i^l close to 1), will influence the decisions of their neighbors to also prefer this label. As a consequence, the label confidence for images where the classifier was uncertain is increased, and even labeling errors can be corrected.

Formally, the label propagation step itself consists of solving a quadratic optimization problem.

$$\min_{(f_1^l,\ldots,f_n^l)\in\mathbb{R}^n} \sum_{i\in L}(f_i^l - y_i^l)^2 + \frac{\lambda}{2}\sum_{i,j\in X} w_{ij}(f_i^l - f_j^l)^2 \quad (2)$$

where the first term is a loss term, and the second term is a (Laplacian) regularizer [1]. A closed form expression for the solution solution $f_1^l,\ldots,f_n^l$ exists, see [20]. The result are real-valued confidences f_i^l for each label l and image i. We can use these to rank embryo images by their label confidence, for example in an interactive labeling tool, or make a hard assignment to the label of maximum response:

$$l_i^* = \arg\max_l f_i^l$$

thereby achieving a fully automatic setup.

3. Experiments

We performed experiments on synthetic data and on Drosophila embryo gene expression images.[1] In both experiments, we use a linear SVM as our base classifier. In order to find optimal cost parameter of the model, we use cross-validation and search for it on a grid with exponential spacing. We pick the values for two hyper-parameters of label propagation (bandwidth, and regularization weight) on the validation set in an analogous way.

3.1. Illustrative toy example

We first show the power of label propagation with two orthogonal sources of information in the following synthetic toy example.

We generate data by sampling from a mixture of Gaussian with three components in $\mathbb{R}^2$ (Figure 5, left). We train the base classifier using the first dimension only. Because the Gaussians overlap strongly in this representation, a large fraction of the data is misclassified (Figure 5, middle).

The second dimension, we use to compute similarities between samples and form the respective neighborhood graph. Running label propagation many of the previously misclassified examples are now assigned to the correct class

[1]Both data and code is available from http://cvml.ist.ac.at/drosophila/

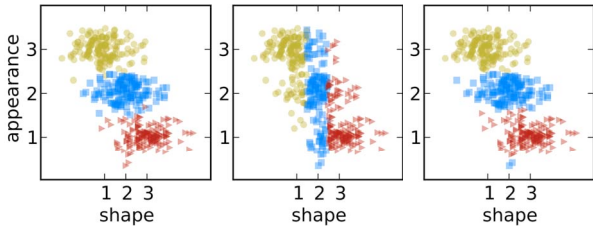

Figure 5. Synthetic test data (left) and corresponding results from the base classifier (middle) and label propagation (right).

(Figure 5, right). The reason is that samples tend to have more neighbors of their own class than of other classes. Since correctly classified examples typically have a higher confidence score, these samples' scores are pushed towards their correct classes. Note that here we think of the second dimension as similarity — only the relative distances matter, ie. directly applying approaches such as PCA is not possible. This is also why some outliers remain mislabeled in Figure 5.

3.2. Drosophila embryos stage annotation

Next, we describe the actual Drosophila stage annotation system. We describe how to obtain the data, and we evaluate the performance of stage annotation by per-embryo shape-based classification and show how it improves by label propagation.

3.2.1 Training data collection

We perform stage-specific collections of embryos as described in Section 2.2 for a Drosophila line serving as the background genotype for the study and therefore showing no expression patterns. We exclude abnormal embryos (blurred, overlapping, broken, incorrect stage) yielding a total of 6810 embryos along with their contours from 47 slides. We explicitly balance this dataset resulting in 1014 embryos per stage group.

3.2.2 Validation and test data collection

We collected expert annotations for 193 slides from the described biological study (embryos show various expression patterns). Two annotators were given a set of automatically preselected embryos for each of the slides separately. At least a hundred embryos were selected with emphasis on non-blurred, non-overlapping embryos showing any active pattern. If there were not sufficiently many such automatically selected embryos the set was completed by random sampling. The annotators were free in which embryos to label, with the instruction to try to annotate at least one embryo of each stage that shows any active pattern.

The first expert annotated 49 slides with 102 embryos annotated on average, the second expert annotated 150 slides with 27 embryos per slide, in total we obtained 9006 annotated embryos. As there are usually more embryos of later stages and the activity is more frequent in later stages as well, the dataset is significantly unbalanced in favor of later stages. The stage groups 4–6, 7–8, 9–10, 11–12, 13–14, and 15–16 are represented by 205, 176, 802, 1957, 2678, and 3188 embryos.

We randomly select 20 slides as a validation set for tuning the parameters of label propagation. The remaining 173 serve as test set. These were not used during any part of the method development, but serve only to estimate the generalization performance of the system on future data. The test set contains 6 slides which were annotated independently by both experts. Each of these slides is included in the test set just once. We use these duplicates to get a rough estimate of the human error rate on the task.

3.2.3 Embryo stage annotation

We choose the parameters of the shape features based on prior experience on a smaller dataset. We use chain codes of length 8, we consider chain codes with Hamming distance 2 or less to be adjacent and we take 5000 chain codes having highest term frequency-inverse document frequency for the purpose of building the dictionaries. We use three scales of the contour, extracted from images at resolution 800×400px and smaller subsampled by factors of 2 and 4, and finally we take the square root of the feature entries, corresponding to a Hellinger kernel feature map.

To evaluate the quality of the SVM classifier, we perform five different runs based on different subsets of the training data and apply the resulting classifiers to the test set. We measure the quality of prediction by the *slide accuracies*, *i.e*. the average number of correct predictions per slide, averaged over all slides of the test set. The results are shown in Table 1. Furthermore, Figure 6 (left) depicts the confusion matrix for one of the runs.

Most of the classification errors are caused by mistaking two neighboring stage groups. We can see that the accuracy for some of the stages is reasonably good, especially 4–6 which has very smooth contour, 13–14 where there are many deep ridges between the segments, and 15–16 where the segment boundaries get slightly softer. On the other hand, the stage 9–10 is very hard to predict resulting in less than 40% accuracy. This is due to the contours often lacking characteristic features that would help to distinguish this stage group.

Next, we obtained the labels from a classifier on the validation set, and use it to identify optimal parameters (bandwidth, regularization weight) of the label propagation. With these parameters fixed we run label propagation on the test set. For all of the runs the slide accuracy improves the results by about 10%, see Table 1 and Figure 6 (right).

run	slide accuracy classifier	slide accuracy propagation
A	0.721	**0.814**
B	0.719	**0.818**
C	0.721	**0.814**
D	0.724	**0.815**
E	0.698	**0.809**

Table 1. Comparison of classifier and label propagation on the test set measured by average slide accuracy. Label propagation consistently brings improvement of about 10%.

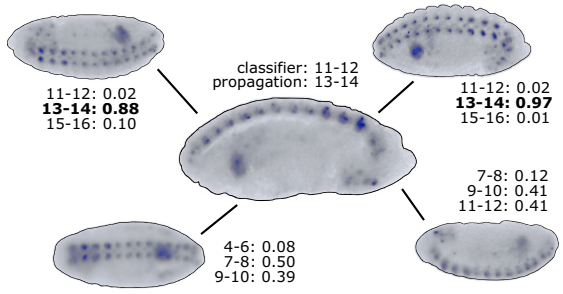

Figure 7. A late stage 14 embryo incorrectly predicted by a classifier as 11–12 gets the right label upon label propagation thanks to some of its nearest neighbors being confidently and correctly predicted (top right, top left). Part of the marginal from classifier output corresponding to the three most probable stages are shown. Contours of embryos marked for easier interpretation.

In Figure 7, we show one particular example where the contour of an embryo does not reflect well the correct stage and the embryo is labelled incorrectly by the SVM classifier. This error is corrected by label propagation since there are enough embryos which are confidently predicted to be of the correct stage and which have very similar pattern.

The improvement of label propagation as measured by average slide accuracy is largely due to later stages (11–12, 13–14, 15–16) which contribute to this improvement the most, see Figure 6 (right).

However for some applications, other accuracy measures might be more suitable. We report *sample accuracy*, *i.e*. the fraction of embryos correctly classified across all slides, and *label accuracy*, *i.e*. the fraction of correctly classified embryos within each stage, averaged over all stages. Also using these measures label propagation improves over SVM classification, see Table 2. Overall, slide accuracy and sample accuracy have higher values compared to label accuracy because the majority of the embryos belongs to the more improved later stages.

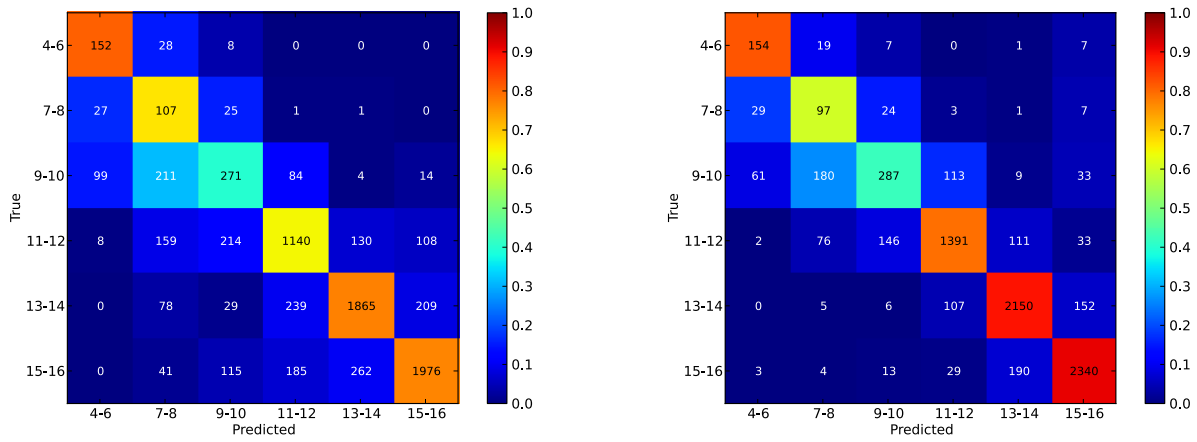

Figure 6. A confusion matrix for run A on the test set after SVM prediction only (left), and after label propagation (right). The numbers indicate embryo counts, the colors encode per-class fractions.

run	sample accuracy		label accuracy	
	classif.	propag.	classif.	propag.
A	0.707	**0.824**	0.679	**0.692**
B	0.709	**0.828**	0.683	**0.695**
C	0.709	**0.825**	0.686	**0.700**
D	0.711	**0.823**	0.679	**0.694**
E	0.687	**0.820**	0.681	**0.706**
human	0.83		0.62	

Table 2. Sample and average label accuracies of classifier and label propagation on the test set. In all cases, label propagation improves over shape-based SVM classification, and it is well comparable to human annotators.

3.2.4 Discussion

Overall, label propagation improves over the classification in all sample, label, and slide accuracies. From the two confusion matrices (Figure 6), we can see, however, that there are two labels which are challenging for our system. First, stage group 9–10 is difficult to classify based on contour and leads to common mistakes, label propagation then corrects these mistakes only sporadically. This happens most probably because there are not enough correct predictions of this stage. Second, stage group 7–8 is the only label for which label propagation leads to a decrease in accuracy. This stage group is unique in the sense that it lasts only 35 minutes (compared to 90–260 minutes for the other five stage groups) which results in very few embryos collected and mounted on one particular slide. Stage group 7–8 is also the point when the embryo undergoes highly dynamic changes (final gastrulation and rapid germ band elongation [3]) which result in diverse expression patterns. For this stage one cannot expect a significant improvement from the label propagation. If a 7–8 embryo is labeled incorrectly by the SVM classifier, its decision will likely not be corrected, because none of the other embryos of this stage have a similar enough pattern to play a role during propagation.

4. Conclusion

We proposed a system for Drosophila embryo stage classification that combines a classifier and graph-based label propagation. The classifier is rather weak but can be learned from general shape features, whereas the graph captures specific appearance of each of the stages for one particular genotype and enables us to correct for the errors of the classifier. This setup minimizes the need for the training data and still gives a system with performance comparable to human annotators.

One important advantage of the presented system is that it can also be easily used in a semi-automatic way, which many biologists prefer over fully automatic systems. First, the label propagation outputs real-valued scores that can be used for ranking instead of a hard decision. Second, it is possible to include input from the user in the form of hard labels without any significant changes to the algorithm. Only the label propagation has to be rerun, which takes minimal computation effort. A further advantage of the semi-automatic scenario is that the user can change the parameters of label propagation on the fly, thereby avoiding the need for model selection on a validation set.

The main limitation of the proposed system is that its performance depends strongly on the quality of the neighborhood graph, i.e. on the similarity measure. Also, rare classes can get suppressed by frequent ones in the label propagation step as we saw with the early developmental stages which are very short.

In future work, we plan to improve the classification accuracy specifically for the rare and difficult stages. Apart from better imaging techniques, possible directions for this include the development of better shape and appearance features, as well as non-linear classifiers. Note that since our method is modular, we can analyze and improve each of those aspects in isolation to find the combination with an overall best performance.

We plan to use the system to automatically prefilter the millions of individual embryo images and identify a small set of embryo images per stage, from which a human picks the visually most suitable one for biological interpretation.

Acknowledgement

We would like to thank Katharina Schernhuber and Michaela Pagani for their help with the embryo collections. This work was in parts funded by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreements nos 242922 and 308036. Basic research at the IMP is supported by Boehringer Ingelheim GmbH.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research (JMLR)*, 7:2399–2434, 2006.

[2] X. Cai, H. Wang, H. Huang, and C. Ding. Joint stage recognition and anatomical annotation of Drosophila gene expression patterns. *Bioinformatics*, 28(12):116–124, 2012.

[3] J. A. Campos-Ortega and V. Hartenstein. *The embryonic development of Drosophila melanogaster*. Springer, 1997.

[4] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[5] H. Freeman. On the encoding of arbitrary geometric configurations. *IRE Transactions on Electronic Computers*, EC-10(2):260–268, 1961.

[6] E. Frise, A. S. Hammonds, and S. E. Celniker. Systematic image-driven analysis of the spatial Drosophila embryonic expression landscape. *Molecular Systems Biology*, 6:345, 2010.

[7] S. Ji, L. Yuan, Y.-X. Li, Z.-H. Zhou, S. Kumar, and J. Ye. Drosophila gene expression pattern annotation using sparse features and term-term interactions. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[8] S. Kumar, K. Jayaraman, S. Panchanathan, R. Gurunathan, A. Marti-Subirana, and S. J. Newfeld. BEST: a novel computational approach for comparing gene expression patterns from early stages of Drosophila melanogaster development. *Genetics*, 162(4):2037–2047, 2002.

[9] E. Z. Kvon, T. Kazmar, G. Stampfel, J. O. Yáñez-Cuna, M. Pagani, K. Schernhuber, B. J. Dickson, and A. Stark. Genome-scale functional characterization of developmental enhancers in Drosophila. (submitted).

[10] C. H. Lampert. Kernel methods in computer vision. *Foundations and Trend in Computer Graphics and Vision*, 4:193–285, 2009.

[11] Y.-X. Li, S. Ji, S. Kumar, J. Ye, and Z.-H. Zhou. Drosophila gene expression pattern annotation through multi-instance multi-label learning. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, 2009.

[12] H. Peng and E. W. Myers. Comparing in situ mRNA expression patterns of Drosophila embryos. In *International Conference on Resaerch in Computational Molecular Biology (RECOMB)*, 2004.

[13] B. D. Pfeiffer, A. Jenett, A. S. Hammonds, T.-T. B. Ngo, S. Misra, C. Murphy, A. Scully, J. W. Carlson, K. H. Wan, T. R. Laverty, C. Mungall, R. Svirskas, J. T. Kadonaga, C. Q. Doe, M. B. Eisen, S. E. Celniker, and G. M. Rubin. Tools for neuroanatomy and neurogenetics in Drosophila. *Proceedings of the National Academy of Sciences*, 105(28):9715–9720, 2008.

[14] K. Puniyani, C. Faloutsos, and E. P. Xing. SPEX^2: automated concise extraction of spatial gene expression patterns from fly embryo ISH images. *Bioinformatics*, 26(12):147–156, June 2010.

[15] K. Puniyani and E. P. Xing. Inferring gene interaction networks from ISH images via kernelized graphical models. In *European Conference on Computer Vision (ECCV)*. Springer, 2012.

[16] P. Tomancak, A. Beaton, R. Weiszmann, E. Kwan, S. Shu, S. E. Lewis, S. Richards, M. Ashburner, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 3(12):81–88, 2002.

[17] P. Tomancak, B. P. Berman, A. Beaton, R. Weiszmann, E. Kwan, V. Hartenstein, S. E. Celniker, and G. M. Rubin. Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, 8(7):R145, 2007.

[18] J. Ye, J. Chen, Q. Li, and S. Kumar. Classification of Drosophila embryonic developmental stage range based on gene expression pattern images. In *Computational System Bioinformatics Conference*, 2006.

[19] J. Zhou and H. Peng. Automatic recognition and annotation of gene expression patterns of fly embryos. *Bioinformatics*, 23(5), 2007.

[20] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning (ICML)*, 2003.